

# Inhalt

<b>1</b>	<b>Einleitung</b> .....	11
1.1	Thematischer Hintergrund .....	11
1.2	Terminologie und Definitionen .....	12
1.3	Zielsetzung .....	13
1.4	Thematische Eingrenzung .....	14
1.5	Methodische Vorgangsweise .....	15
1.6	Gliederung der Arbeit .....	16
<b>2</b>	<b>Zur Methodik des automatischen Klassifizierens</b> .....	17
2.1	Automatisches Klassifizieren von Textdokumenten .....	17
2.1.1	Begriffsumfang .....	17
2.1.2	Einfach- und Mehrfachklassifizierung .....	19
2.1.3	Klassen- vs. Dokumentenzentrierung .....	19
2.1.4	"Harte" vs. rangordnende Klassifizierung .....	20
2.2	Hauptanwendungen der automatischen Textklassifizierung .....	20
2.3	Maschinelle Lernverfahren .....	21
2.3.1	Trainings-, Test- und Validierungsdokumente .....	22
2.3.2	Information-Retrieval-Techniken und Textklassifizierung .....	23
2.4	Dokumentenindexierung und Merkmalsreduktion .....	23
2.4.1	Indexierung der Dokumente .....	23
2.4.2	Dimensionsreduktion .....	24
2.5	Induktive Erstellung von Klassifikatoren .....	26
2.5.1	Bestimmung von Schwellenwerten .....	27
2.5.2	Arten von Klassifikatoren .....	27
2.5.3	Kombination von Klassifikatoren .....	31
2.6	Evaluierung der Klassifizierungsgüte .....	31
2.6.1	Masse für die Klassifizierungsgüte .....	31
2.6.2	Benchmarks .....	34
2.6.3	Suche nach dem besten Klassifikator .....	35
2.7	Labor-, Open Source und kommerzielle Software .....	36

<b>3</b>	<b>Die Projekte an der Universität Lund</b>	39
3.1	Nordic WAIS / WWW	39
3.1.1	Methodische Vorgangsweise	39
3.1.2	Evaluierung	40
3.1.3	Benutzung	40
3.2	DESIRE II	41
3.2.1	Engineering Electronic Library, Sweden, und "All" Engineering	42
3.2.2	Ei-Klassifikation und Ei-Thesaurus	43
3.2.3	Klassifizierungsprozess	44
3.2.4	Evaluierung	48
3.2.5	Benutzung	50
3.2.6	Anwendung anderer Klassifizierungsverfahren	50
3.2.7	Thematisches Vorfiltern beim Web-Harvesting	51
3.2.8	Exkurs: SOSIG	51
3.3	Engine-e	51
3.3.1	Methodische Vorgangsweise	52
3.3.2	Evaluierung	52
3.3.3	Benutzung	53
<b>4</b>	<b>Wolverhampton Web Library (The UK Web Library)</b>	55
4.1	WWLib-TOS und "Old ACE"	55
4.1.1	Aufbereitung des DDC-Vokabulars	55
4.1.2	Klassifizierungsprozess	56
4.1.3	Evaluierung	59
4.1.4	Benutzung	60
4.2	WWLib-TNG und ACE	60
4.2.1	Klassifizierungsprozess	62
4.2.2	Evaluierung	63
4.3	Weitere Experimente mit ACE	64
4.3.1	Adaptives automatisches Klassifizieren mit ACE	64
4.3.2	Ontologie-basiertes automatisches Klassifizieren mit ACE	65
<b>5</b>	<b>German Harvest Automated Retrieval and Directory</b>	67
5.1	Das DFG-Projekt GERHARD	67
5.1.1	UDK und UDK-Lexikon	68
5.1.2	Klassifizierungsprozess	70

5.1.3	Evaluierung .....	72
5.1.4	Benutzung .....	73
5.2	GERHARD und DESIRE II .....	74
5.3	Das Nachfolgeprojekt GERHARD II .....	75
5.3.1	Intentionen .....	75
5.3.2	Gegenwärtiger Entwicklungsstand .....	77
<b>6</b>	<b>Das Projekt Scorpion von OCLC .....</b>	<b>79</b>
6.1	Überblick .....	79
6.2	Die Dewey Datenbank .....	81
6.2.1	Varianten der Dewey-Datenbank .....	81
6.2.2	Vokabularanreicherung .....	83
6.2.3	Test der DDC auf Klassenintegrität .....	85
6.3	Behandlung der Input-Dokumente .....	87
6.4	Klassifizierungsverfahren .....	87
6.5	Nachbearbeitung der Ergebnisse .....	88
6.6	Evaluierung .....	90
6.6.1	Masse für den Vergleich von DDC-Ergebnismengen .....	90
6.6.2	Die NetFirst-Studie .....	91
6.7	Scorpion und DESIRE II .....	93
6.8	Scorpion und die LCC .....	94
6.9	Benutzungsmöglichkeiten .....	96
6.9.1	CORC / Connexion .....	96
6.9.2	WWW-Klassifikatoren .....	97
6.9.3	Open Source Version .....	97
<b>7</b>	<b>Weitere Anwendungen und Projekte .....</b>	<b>99</b>
7.1	Automatisches Klassifizieren von Büchern .....	99
7.1.1	Die LCC-Studie von Larson .....	99
7.1.2	Das ACS-Verfahren von Cheng & Wu .....	102
7.1.3	Sonstige Anwendungen und Projekte .....	103
7.2	Automatisches Klassifizieren von Patentliteratur .....	105
7.2.1	Tests des U.S. Patentamtes .....	106
7.2.2	Tests des Europäischen Patentamtes .....	108
7.2.3	Tests der WIPO.....	109

7.2.4	Die französische IPC-Suchmaschine .....	111
7.2.5	Das japanische Klassifizierungssystem OWAKE .....	114
7.3	Automatisches Klassifizieren in der Mediendokumentation .....	116
7.3.1	Gruner + Jahr .....	116
7.3.2	Zweites Deutsches Fernsehen .....	117
7.3.3	Bayerischer Rundfunk und Süddeutscher Verlag .....	118
7.3.4	Artikel aus belgischen Magazinen .....	119
7.3.5	Andere Untersuchungen an Presstexten .....	119
7.4	Einsatz bei Web-Portalen, Suchmaschinen, Informationsdiensten .....	120
7.4.1	Lexis-Nexis .....	120
7.4.2	Northern Light .....	121
7.4.3	Factiva .....	123
7.4.4	INFOMINE .....	125
7.4.5	Sonstige Anwendungen und Projekte .....	126
<b>8</b>	<b>Diskussion und Ausblick .....</b>	<b>133</b>
8.1	Zur Methodik des automatischen Klassifizierens .....	133
8.2	Die Projekte an der Universität Lund .....	134
8.3	Wolverhampton Web Library .....	135
8.4	GERHARD .....	137
8.5	Scorpion / OCLC .....	138
8.6	Weitere Anwendungen und Projekte .....	139
8.7	Andere Aspekte .....	141
8.8	Ausblick .....	145
<b>9</b>	<b>Literaturverzeichnis .....</b>	<b>149</b>
	<b>Anhang .....</b>	<b>173</b>
	Verwendete Abkürzungen und Akronyme .....	173
	Abbildungs- und Tabellenverzeichnis .....	176